COMPSCI 630: Randomized Algorithms	January 24, 2018
Lecture 3	
Lecturer: Rong Ge	*Scribe: Alex Steiger*

1 Overview

In this lecture, we will introduce the common model of *balls and bins* and analyze a number of example problems using it. The setting is that we have n bins, and balls are thrown independently and uniformly at random into them, and thus any ball ends up in any given bin with probability $1/n$. We can ask a number of questions in this setting, such as:

1. How many balls should be thrown before a bin has two balls in it?
2. How many balls should be thrown before every bin has at least one ball in it?
3. After throwing m balls, how many balls are in the bin with the most balls?

In today's lecture, we will focus on two problems which we will formulate in this setting, the ***birthday paradox*** and ***coupon collector problem***.

2 Preliminaries

First, we recall a few inequalities that we will use.

Definition 1 (Markov's inequality)**.** *Let X be a non-negative random variable. Then for any real number $t > 0$, $\Pr[X \geq t] \leq \mathrm{E}[X]/t$.*

Definition 2 (Chebychev's inequality)**.** *Let X be a non-negative random variable. Then for any real number $t > 0$, $\Pr[|X - \mathrm{E}[X]| \geq t] \leq \mathrm{Var}[X]/t^2$.*

Definition 3 (Union bound)**.** *Let $\{A_1, A_2, \ldots, A_n\}$ be a set of events. Then* $\Pr\left[\bigcup_{1 \leq i \leq n} A_i\right] \leq \sum_{1 \leq i \leq n} \Pr[A_i]$.

3 The birthday paradox

Consider a room of m people. It is reasonable to consider that their m birthdays are independent and each are any given day of the year with equal probability. What is the probability that two people have the same birthday? By the pigeonhole principle, if $m > 365$, then this occurs with probability 1. The ***birthday paradox*** states that when $n = 23$, the probability is roughly $1/2$. In this section we will try to answer this question after formulating it in the balls and bins setting: after throwing m balls into n bins, what is the probability that a bin has at least two balls in it? By setting $n = 365$, this is equivalent to the original question set out (ignoring leap days).

3.1 Idea 1: Compute the exact probability using conditioning

First, we will attempt to compute the exact probability as follows. For $1 \leq i \leq m$, let A_i be the event that after throwing the first i balls, all lie in different bins. Then we have

$$\begin{aligned}
\Pr[A_{i+1}] &= \Pr[A_{i+1} \mid A_i] \cdot \Pr[A_i] + \Pr[A_{i+1} \mid \neg A_i] \cdot \Pr[\neg A_i] && (1)\\
&= \Pr[A_{i+1} \mid A_i] \cdot \Pr[A_i] + 0 && (2)\\
&= \Pr[\,i\text{'th ball thrown into an empty bin} \mid A_i] \cdot \Pr[A_i] && (3)\\
&= \frac{n-i}{n} \cdot \Pr[A_i] && (4)\\
&= \left(1 - \frac{i}{n}\right) \cdot \Pr[A_i]. && (5)
\end{aligned}$$

Line 1 follows from the law of total probability, line 2 follows from the fact that $\Pr[A_{i+1} \mid \neg A_i] = 0$, and line 3 follows by definition of A_{i+1}. Then line 4 follows since A_i implies exactly $n-i$ bins are empty, and the $(i+1)$'th ball ends up in any bin with equal probability. Since $\Pr[A_1] = 1$, it follows by induction that:

$$\Pr[A_m] = \left(1 - \frac{m-1}{n}\right)\left(1 - \frac{m-2}{n}\right)\left(1 - \frac{m-2}{n}\right)\ldots\left(1 - \frac{2}{n}\right)\left(1 - \frac{1}{n}\right) \tag{6}$$

First, note that this exact probability is similar to the success probability of Karger's global min-cut algorithm from the first lecture, but not as nice to simplify. Here, we employ our first trick to lower bound this probability: for any $a, b \in [0,1]$, $(1-a)(1-b) = 1-a-b+ab \geq 1-a-b$. (Note the similarity of this inequality to the union bound.) Repeatedly applying this inequality, we have:

$$\Pr[A_m] \geq 1 - \frac{m-1}{n} - \frac{m-2}{n} - \cdots - \frac{2}{n} - \frac{1}{n} \geq 1 - \frac{m(m-1)}{2n} \tag{7}$$

This implies that when m is sufficiently small (say a small constant), probability is close to 1; however, for $m \approx \sqrt{2n}$, we get a lower bound close to zero, which is arguably useless since the event clearly occurs with some positive probability.

Using another trick, we will obtain an upper bound on $\Pr[A_m]$. First, recall the series expansion $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots$, then note $1 + x < e^x$. In particular, $1 + x$ is a decent approximation to e^x for $x \in [-1, 1]$. It follows that:

$$\Pr[A_m] \leq \exp\left(-\frac{m-1}{n}\right)\exp\left(-\frac{m-2}{n}\right)\ldots\exp\left(-\frac{2}{n}\right)\exp\left(-\frac{1}{n}\right) = \exp\left(-\frac{m(m-1)}{2n}\right) \tag{8}$$

where $\exp(x) = e^x$. Then, from this upper bound and our previous lower bound, we have that when $m \approx \sqrt{(2-\varepsilon)n}$ for a constant $\varepsilon > 0$, we have $\Pr[A_m] = \Theta(1)$.

3.2 Idea 2: Use the union bound and linearity of expectation

For $1 \leq i < j \leq m$, let $B_{i,j}$ be the event that the i'th ball and j'th ball are thrown into the same bin. Clearly $\Pr[B_{i,j}] = 1/n$. By definition, we have:

$$A_m = \bigcap_{1 \leq i < j \leq n} \neg B_{i,j} \tag{9}$$

$$\implies \neg A_m = \bigcup_{1 \leq i < j \leq n} B_{i,j}. \tag{10}$$

By applying the union bound to $\neg A_m$, we have

$$\Pr[\neg A_m] \leq \sum_{1 \leq i < j \leq m} \Pr[B_{i,j}] = \frac{m(m-1)}{n} \tag{11}$$

which gives us the same lower bound of $1 - \frac{m(m-1)}{2n}$ we previously obtained for $\Pr[A_m]$.

Alternatively, we can get the same bound again by computing the expected number of *collisions*, i.e. the number of times a distinct pair of balls are thrown into the same bin, then applying Markov's inequality to obtain a lower bound on $\Pr[A_m]$. Note that the number of collisions is equal to the number of $B_{i,j}$'s that occur, and A_m occurs if and only if that number is zero. Then we have

$$\mathrm{E}\left[\sum_{1 \leq i < j \leq m} \mathbb{1}_{A_{i,j}}\right] = \sum_{1 \leq i < j \leq m} \mathrm{E}\left[\mathbb{1}_{A_{i,j}}\right] = \sum_{1 \leq i < j \leq m} \Pr[A_{i,j}] = \frac{m(m-1)}{2n} \tag{12}$$

where $\mathbb{1}_x$ is the indicator variable for an event x. By Markov's inequality, it follows that:

$$\Pr[\neg A_m] = \Pr\left[\sum_{1 \leq i < j \leq m} \mathbb{1}_{B_{i,j}} \geq 1\right] \leq \frac{m(m-1)}{2n}. \tag{13}$$

4 The coupon collector problem

In the ***coupon collector problem***, there are n different types of coupons, and you want to collect (at least) one of each type. You obtain the coupons one at a time, and each coupon's type is sampled independently and uniformly at random. In expectation, how many coupons will you obtain before you have one of each type? To answer this question, we will formulate as throwing balls into bins: in expectation, how many balls will be thrown before all n bins are non-empty?

4.1 Idea 1: Compute the expectation directly

Let T be the number of balls thrown so that all bins are non-empty, and let T_i be the number of balls thrown to go from $i-1$ non-empty bins to i non-empty bins. Then $T = \sum_{i=1}^{n} T_i$. For example, let $m = 5$. Consider the following sequence of indices which denote the bins that the balls were thrown into:

$$3 \mid 3,2 \mid 1 \mid 2,3,1,5 \mid 1,1,4 \mid \tag{14}$$

The vertical bars are placed immediately after the first occurrences of each index so that the i'th bar denotes the end of the i'th phase. That is, the first ball landed in bin 3, ending the first phase and starting the second.

Then the second ball also landed in bin 3, followed by the third ball landing in bin 2, ending the second phase and starting the third, and so on. In this example, $\langle T_1, T_2, T_3, T_4, T_5 \rangle = \langle 1, 2, 1, 4, 3 \rangle$, and thus $T = 11$.

By linearity of expectation, we have $\mathrm{E}[T] = \sum_{i=1}^{n} \mathrm{E}[T_i]$, so it is sufficient to compute $\mathrm{E}[T_i]$ for each i. At this point, we recognize T_i is a geometric random variable. That is, with probability $p = (n-(i-1))/n$, the i'th phase ends by the ball being thrown into one of the $n-(i-1)$ empty bins, or the phase continues with probability $1-p$ by the ball being thrown into one of the $(i-1)$ non-empty bins. In other words, for any integer $k > 0$, we have

$$\Pr[T_i = k] = \left(\frac{i-1}{n}\right)^{k-1} \left(\frac{n-(i-1)}{n}\right) \tag{15}$$

and thus $\mathrm{E}[T_i] = 1/p = n/(n-(i-1))$. It follows that:

$$\begin{aligned}
\mathrm{E}[T] &= \sum_{i=1}^{n} \frac{n}{n-(i-1)} \\
&= n \cdot \left(\frac{1}{n} + \frac{1}{n-1} + \ldots \frac{1}{2} + 1\right) \\
&= n \cdot H_n \\
\implies n \log n \leq \mathrm{E}[T] &\leq n \log(n+1) + n \\
\implies E[T] &= \Theta(n \log n).
\end{aligned}$$

Note that the implication follows from the fact that the n'th harmonic number, H_n, is bounded between $\log n$ and $\log(n+1)+1$ (where log is the natural logarithm). Therefore, we should expect to obtain roughly $n \log n$ coupons before we collect one of each type. However, without further analysis, this does not give any insight on how likely it is for us to collect more or less than the expected number. To gain this insight, we will apply Chebychev's inequality (as defined in the preliminaries section), which requires us to upper bound the variance of T, denoted by $\mathrm{Var}[T]$.

Recall that for independent random variables X and Y, $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$. Clearly T_i and T_j for $i \neq j$ are independent; loosely speaking, the length of one phase does not affect the length of another since all ball throws are independent. Furthermore, since T_i is a geometric random variable with success probability $p = (n-(i-1))/n$, we know $\mathrm{Var}[T_i] = (1-p)/p^2$. It follows that

$$\begin{aligned}
\mathrm{Var}[T] &= \mathrm{Var}\left[\sum_{i=1}^{n} T_i\right] \\
&= \sum_{i=1}^{n} \mathrm{Var}[T_i] \\
&= \sum_{i=1}^{n} \frac{n(i-1)}{(n-(i-1))^2} \\
&= n^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + \ldots + \frac{1}{2^2} + 1\right) \\
&< \frac{\pi}{6} n^2
\end{aligned}$$

where the last inequality follows from the series expansion of $\pi/6$. Thus, for any constant $c > 0$, we obtain the upper bound

$$\Pr[|T - \mathrm{E}[T]| \geq cn] \leq \frac{\mathrm{Var}[T]}{c^2 n^2} < \frac{1}{c^2} \tag{16}$$

by Chebyshev's inequality.

4.2 Idea 2: Use the union bound

Instead of computing the expected number of balls thrown before all bins are non-empty, we can try to compute the probability of successfully ending with all non-empty bins after a fixed number of throws, m. For $1 \leq i \leq m$, let A_i be the event that the i'th bin is empty after all m throws. Then we have success if and only if no event A_i occurs for each i. The probability of all m (independent) throws avoid the i'th bin is $(n-1)/n = (1-1/n)$, so we have:

$$\Pr[A_i] = (1-1/n)^m = ((1-1/n)^n)^{m/n} \approx e^{-m/n}. \tag{17}$$

By applying the union bound we get:

$$\Pr[\text{failure}] = \Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \Pr[A_i] = ne^{-m/n}. \tag{18}$$

It follows that when $m = n\log n + n$, the probability of failure at most $ne^{-(n\log n+n)/n} = ne^{-\log n-1} = e^{-1}$, so the probability of success is at least $1-1/e$ for that case.

5 Summary

In this lecture, we went through some simple problems in the setting of independently throwing balls into bins. We used Markov's inequality, Chebyshev's inequality, and the union bound in order to obtain upper and lower bounds for various probabilities and expectations. We also saw the trick of approximating $1+x$ with e^x for $x \in [-1,1]$ which can be quite useful when dealing with probabilities.